J. Environ. Nanotechnol.
Volume 6, No.1(2017) pp. 51-54
ISSN (Print): 2279-0748
ISSN (Online): 2319-5541
doi:10.13074/jent.2017.03.171225

Performance Analysis of Optimization Algorithms in Multiple Nucleotide Sequence Alignment Problem

J. Priyadharshini[1], S. P. Victor[2*]

[1]*Department of Computer Science and Applications, St. Joseph's College for Women, Tirupur, TN, India.*
[*2]*Head & Director of the Research Centre, Department of Computer Science, St. Xavier's College, Palayamkottai, TN, India.*

Abstract

The ultimate aim of the survey is to outline the characters of popular evolutionary algorithms and compare their performances to solve multiple nucleotide sequence alignment problem in bioinformatics. Bioinformatics is the storage, manipulation and analysis of biological information such as nucleic acid and protein sequences via computer science. Multiple sequence alignment is used to generate a concise, information rich summary of sequence data in order to make decisions on the relatedness of sequences to a gene family. Optimization algorithms, such as the Genetic Algorithm(GA),Particle Swarm Optimization (PSO) algorithm, Ant Colony Optimization(ACO)algorithm and Artificial Bee Colony (ABC)algorithm, can give solutions to multiple nucleotide sequence alignment problems near to the optimum for many applications; however, in some cases, they can su□er from becoming trapped in local optima. The Stem Cells Algorithm (SCA) is an optimization algorithm inspired by the natural behavior of stem cells in evolving themselves into new and improved cells. The SCA avoids the local optima problem successfully. Multiple Sequence alignment provides an effective way to find conserved regulatory patterns in nucleotide sequences which helps in the diagnosis and classification of diseases. Since multiple sequence alignment is an ongoing research area, we intend to analyze and compare the features of optimization algorithm which have its own strengths and weaknesses. The proposed hybrid approach of genetic algorithm with an alignment improver of stem cells algorithm produce better results than other optimization algorithms.

Keywords *: Chromosomes; Crossover; Genes; Genetic algorithm; MSA; Mutation; Stem cell.*

1. INTRODUCTION

Sequence similarity plays a major role in Bioinformatics and molecular biology. This stems from the widely accepted conjecture in Molecular Biology that proteins or genes that have similar sequences are likely to perform the same function. One of the most widely used techniques for sequence comparison is sequence alignment. Sequence alignment allows mismatches and insertion/deletion**,** which represents biological mutations. Sequence alignment is normally performed on two sequences. Multiple sequence alignment, is a natural extension of two-sequence alignment (Yongtao Y et al. 2013). In multiple sequence alignment, the emphasis is likely to find optimal alignment for a group of sequences. DNA stores all genetic information. DNA molecules are chains of nucleotides. There are four different types of nucleotides (Jonathan Shapiro, 2001), denoted by A, T, G, C . Therefore DNA molecules can be represented as strings of letters from relatively small alphabets. If there is a set of strings where the number of strings is k. So, given k strings S= {S_1, S_2, S_3,……..S_k}, we try to find an optimal alignment for those sequences. The purpose is to find S' ={S'_1, S'_2,……S'_k} in optimal way such that (Mohammad Taherdangkoo et al. 2012); S'_i is an extension of S_i by inserting or padding gaps/spaces., □ i, j : $|S'_i| = |S'_j|$ and ₍€ⱼ sim(S'_i, S'_j) is maximized or €ᵢ€ⱼ cost(S'_i, S'_j) is minimized where sim(X,Y) and cost(X,Y) are some sequence similarity and alignment cost functions defined for sequences X and Y. In MSA, we have to define a scoring scheme for evaluating matching letters, mismatching letters and gap penalties. There are two types of alignment (Amie Judith Radenbaugh, 2008), global and local. In global alignment, attempts are made to detect the best alignment of the entire sequences. In local alignment, the best alignment is constructed for segments of sequences with the highest density of matches, while the rest of the sequences are ignored. In this paper, only global alignment is investigated.

Seq1 : a c - - b c d b Seq2 : - c a d b - d

The scoring function of aligning is V (x,y) where x and y are each single character or space. For any two distinct characters x and y, V (y,y)= +2 and V (x,y) = V (-,y) = V (x,-) = -1. From a scoring scheme we can calculate the scoring function for both sequences i.e 3.(2) + 5.(-1) = 1. If S is a string, then |S| denotes no of characters in S. For example, if S = acbcbd, then |S| = 6 and S|3| = b. If we have strings S and T, an alignment A maps S into strings S' and T' that may contain space characters, where |S'| = |T'|, and the removal of spaces from S' and T'(without changing the order of remaining characters) leaves S and T respectively. The value of alignment A is

$$\sum_{i=1}^{L} \sigma(V(S'[i], T'[i]))$$

where L = |S'| = |T'| .In the example above, if S = acbcdb and T=cadbd, then S'=ac--bcdb and T'= -cadb-d-. The above alignment is called pair-wise alignment.

2. METHODOLOGY

MODULE 1: Initialization :Sample Input Sequence

>MMVHLTPMMKSAVTALWGKVNVDMVGGMAL
GRLLVVYPWTQRFFMSFGDLSTPDAVM

S. P. Victor
email:

>MMGLSDGMWQLVLNVWGKVMADIPGHGQMV
LIRLFKGHPMTLMKFDKFKHLKSMDMMKAS

>ALVMDNNAVAVSFSMMQMALVLKSWAILKKD
SANIALRFFLKIFMVAPS

>MMRPMPMLIRQSWRAVSRSPLMHGTVLFARLF
ALMPDLLPLFQYNCRQFSSPMD

We insert gaps in the input sequence to make the initial population of say 10 alignments.

lmax = 61, corresponding to the longest sequence, length(N)=1.2 * lmax =74, gap1= length-len1 =17, gap2= length- len2 =13, gap3= length-len3 =25, gap4= length-len4 =20

>MMVHLT---PM---MKSAV-T-AL-
WGKVNVDMVGGMALGR--LLV-VYPWTQ-R-
FFMSF-GDLSTPDA--VM

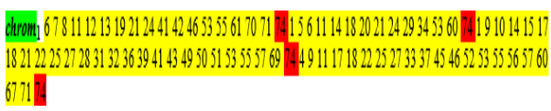
>M-MGL--SDGM-WQ-LVL-N--VW-GKVM-ADIP-
GHGQMVLIRLFKGHPMTL-MKFDKF-KHLKSMD-
MMKAS

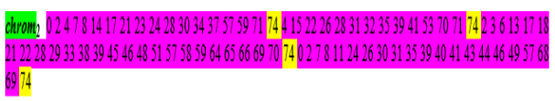
>A-LVMDNNA--VAV--S--FS--MM-Q--MA--LVL-
KS-W-A-ILKKD---S-A-N-IALRFFLKIFM-VAPS

>MMRP-MPML-I-RQSWR--AVS-RS-P-LMHGT-
VLF-ARLFALM--PDLLP--L---FQ-YNCRQF-SSP-MD

MODULE 2: Chromosome Representation

The chromosomes are generated by encoding the sequences. The gap positions in the sequence are being used to represent a chromosome. The gap positions of all the sequences are used to make a single chromosome where, end of a sequence in chromosome representation is indicated by an absolute point. An absolute point's value is equal to the length of each sequence in the initial population. In this manner ten chromosomes are produced corresponding to initial population of 10 alignments.





MODULE 3: Reproduction/Selection

Chromosomes are selected from the population to be parents to crossover and produce offspring.

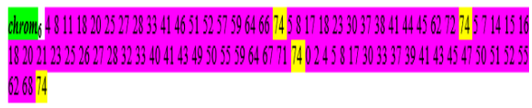
Elitism: In this method, first the best 20% chromosomes are copied to a new population. The rest of the chromosomes undergo genetic operation in a classical manner. Elitism can very rapidly increase the performance of Genetic algorithm because it prevents loosing the best-found solutions. Table 1 shows newly generated 10 chromosomes and their fitness values.

Table 1. Chromosomes and their fitness values

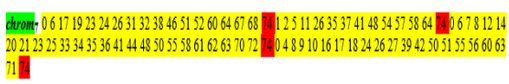
Chrom	Fitness value
1	-597
2	-616
3	-622
4	-637
5	-660
6	-497
7	-694
8	-670
9	-654
10	-616

Applying Elitism- the highest fitness value chromosome, is part of the next generation:

$chrom_6$ Fitness= -497



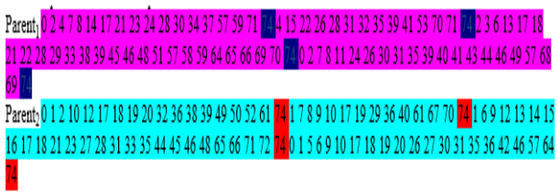
Random Selection: In this method, any random chromosomes are copied to a new population. The rest applying selection

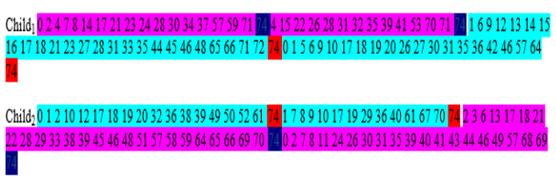


MODULE 4 : Cross Over

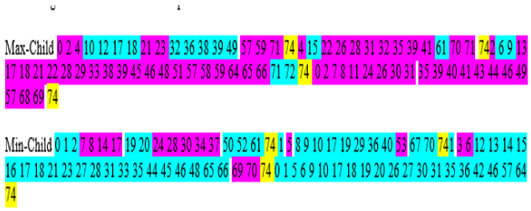
Cross over is a process of taking more than one parent chromosomes and producing a child solution from them. Here, we have implemented three types of crossovers- single point crossover, two point crossover and max-min crossover. Cross over is performed by selecting two parents with higher fitness values as shown in example and then selecting a single crossover point which may be some formula based or randomly determined based on the length of the parents. Each such crossover results in two child chromosomes. As an experimental scheme we have restored only those child chromosomes which have better fitness scores than their parents.

Crossover point = 0.6*74 = 44, nearest absolute point is 32nd position at which crossover is performed.





Performing min-max crossover on the parent chromosomes of above illustration:



MODULE 5 : Mutation

Mutation is a genetic operator used to maintain genetic diversity from one generation of a population of algorithm chromosomes to the next.

Before Mutation:

M--MVHLTPMMK-AVTALWGKVNVDMVG-GMALGR-L-LV--VYP-WTQRFFMS-F-GDLSTPDAVM-GN

After Mutation:

MM-VHLT-PMMKSAV-TALWG-KVNVDMV-GGMALGRLL--VVYP-W-QRFFMSFG---DLSTPDAVMG-N

MODULE 6: Fitness function

In this method, for each location on the aligned sequences, one of three situations will occur: match, mismatch or a gap.

Table 2. Fitness Scores

S1	A	T	-	G	A	T	-	C	C
S2	-	T	A	G	C	T	A	C	C
S3	A	-	A	-	A	T	A	G	C

Fitness Score = Fitness (s1, s2) + Fitness (s1, s3) + Fitness (s3, s2)

Two scoring matrices: PAM-250, BLOSUM-45 [3]

MODULE 7: Proposed alignment improver

Stem Cells Algorithm (SCA), is based on behavior of stem cells in reproducing themselves. SCA has high speed of convergence, low level of complexity with easy implementation process. It also avoids the local minimums in an intelligent manner. The proposed algorithm can be used to solve multiple sequence alignment problems which is expected to rectify the problems faced with the previously analyzed Genetic Algorithm because of its behavioral characteristics. Comparisons can be made with other evolutionary algorithms like particle swarm optimization (PSO) algorithm, ant colony optimization (ACO) algorithm and artificial bee colony (ABC) algorithm in solving the multiple sequence alignment problem.

Two features are considered as the main characteristics in the definition of stem cells that are as follows:

1. Self-renewal: They have the ability of producing from the various cycles of cell division with maintaining the characteristics of that cell.
2. Power: They have the capacity of resolution from different types of specific cells, but it is possible that a cell has also the ability to be separated into several cells. The initial matrix of variables in Stem Cells (Amie Judith Radenbaugh, 2008).

Objective: Improve alignment quality from a single solution produced from GA (if required).

Description:

If the fitness value is less than a threshold value (e.g. <70% fitness), the program will proceed to stem cell algorithm phase to align a solution produced from the GA phase.

From here, stem cell algorithm phase will use the same representation and the same fitness function used in GA phase. One main reason of using Aligning Improver (SCA phase) is to avoid local minima. In the hybrid system of GA/SCA mentioned earlier, it produces average results. This is due to the problem of the progressive method in Dynamic Programming that gets easily trapped in local minima. If this happens then the process cannot be repeated once it is at the middle of the progressive alignment. This can be overcome with the use of SCA.

3. CONCLUSION

The genetic algorithm implements the law of survival of the fittest with a focus on the multiple sequence alignment problem which required more number of iterations to meet the convergence point. Stem cell algorithm algorithm is based on natural behavior of stem cells in reproducing themselves. Self renewal and Power relation are the main features of this algorithm which are used to align the multiple sequences with less number of iterations and with high speed of convergence. We have used a hybrid approach of genetic algorithm for aligning sequences and stem cell algorithm as an alignment improver. In order to reveal valuable evolutionary information, this MSA is carried out with the help of this hybrid approach.

REFERENCES

Amie Judith Radenbaugh, Applications of genetic algorithms in bioinformatics, San Jose State University(2008).

Gen, M. and Cheng, R., Genetic Algorithms and Engineering Optimization. John Wiley & Sons: Canada(2000).

Jonathan shapiro, Genetic algorithms in Machine Learning and its applications, Lecture Notes in Computer Science, Lume., 2049, 146-168(2001).

Mohammad Taherdangkoo, Mahsa Paziresh, Mehran Yazdi and Mohammad Hadi Bagheri, An efficient algorithm for function optimization: modified stem cells algorithm, Central European Journal of Engineering, 3(1), 12-29(2013). doi:10.2478/s13531-012-0047-8

Mohammad Taherdangkoo, Mehran Yazdi, and Mohammad Hadi Bagheri, Stem cells optimization algorithm, ICIC 2011, doi: 10.1007/978-3-642-24553-4_52

Omar, M. F., Salam, R. A., Abdullah, R. and Rashid, N. A., Multiple sequence alignment using optimization algorithms, *Int. J. Comp. Intell.* 1(2), 81-89(2005).

Pankaj Agarwal, Alignment of multiple sequences using GA method, *Int. J. Emer. Technologies Computat. App. Sci.*, (IJETCAS), 4(4), 411-421(2013).

Yongtao Ye, David W. Cheung, Yadong Wang, Siu-Ming Yiu , Tak-Wah Lam, Hing-Fung Ting, GLProbs: Aligning multiple sequences adaptively, Proceedings of International conference on Bioinformatics, computational biology and biomedical informatics, (2013).
doi:10.1145/2506583.2506611